DOCUMENT RESUME

ED 416 672 FL 024 860

AUTHOR	Bauman, John
TITLE	Vocabulary Resources for Material Writers.
PUB DATE	1996-10-00
NOTE	5p.; "From 'The Materials Writers Newsletter' of Japan Association of Language Teachers, October 1996."
PUB TYPE	Reports - Descriptive (141)
JOURNAL CIT	Material Writers Newsletter; v4 n3 Oct 1996
EDRS PRICE	MF01/PC01 Plus Postage.
DESCRIPTORS	*English (Second Language); Foreign Countries; Information Sources; *Instructional Materials; Language Research; *Material Development; *Reference Materials; Second Language Instruction; *Vocabulary
IDENTIFIERS	*Language Corpora

ABSTRACT

Intended for writers of instructional materials for teaching English as a Second Language (ESL), the list describes word lists and language corpora that may be of use in creating, simplifying, and refining vocabulary content in ESL materials. The sources are dated from 1944 to the present, and include a freeware computer program. Some limited comments are made about the utility of the lists/corpora. A brief list of related World Wide Web sites and six print references are provided. A resource containing a discussion of vocabulary instruction is also noted. (MSE)

Vocabulary Resources for Material Writers

From *The Materials Writers Newsletter*
The Newsletter of the Materials Writers' National Special Interest Group
of the Japan Association of Language Teachers
Vol. IV, No. 3, October 1996

John Bauman
Temple University Japan

Material written for ESL students needs to use somewhat simplified vocabulary and structure if it is to be accessible to lower and intermediate level students. In terms of vocabulary, a writer can try to "keep it simple" while writing, but a more rigorous approach is to compare a text with a list of words prepared for this purpose. A variety of lists of words are available, as well as different ways to use them. In this article, I will briefly list and describe some lists. I'll also discuss a program that will analyze a text and give some links for further exploration of this topic on the internet. URLs are given in the "Web Links" section following this article.

Teaching and Learning Vocabulary (Nation 1990) contains a good general discussion of this topic. Nation doesn't hesitate to quantify the issue. His model of an ideal vocabulary teaching sequence starts with the most frequent 2,000 words, which he calls general service vocabulary. Everybody needs to know these words; they make up about 87% of an average written text. After this point, general frequency becomes less useful as a guide to what words to teach. Students are better off studying a list of words specific to their field of interest or need, if one can be found. For the student aiming at English-language higher education, Nation's 800 word University Word List is appropriate. After this, the remaining vocabulary of English is of too little frequency to merit direct study. Skills such as analyzing word parts, context guessing, etc. can be taught.

The number of different words used will depend on the level of the text. Writers of material for ESL learners also have to decide which words to use, or, in a larger sense, to which population of words should they restrict themselves. Here a list becomes necessary. Many have been developed over the years. The following remain relevant.

The General Service List

The General Service List (GSL)(West 1953) is the specific list of 2,000 words that Nation refers to when he writes about the "first 2,000 words." It's based on written texts, it's old, and it's not in frequency order, though frequency numbers are given. The source of the frequency information is even earlier than the publication date, being derived from Thorndike and Lorge (1944). But the list was not compiled based on frequency alone. It was created to be an ideal vocabulary for ESL students to start out with. Through the 1970s, a lot of material, particularly graded readers, was based on this list. Even today, much of this material is sold and used. The GSL is out of print, and somewhat out of favor. The list is available as a component of the Vocabprofile program described below and, in a slightly different form, on my web page.

Thorndike and Lorge

The Teacher's Word Book of 30,000 Words (Thorndike and Lorge, 1944) was created as a resource for elementary and high school teachers in the United States. It is still frequently cited, though computer-produced corpora have largely replaced it as an authority on the frequency of words. For example, it's the source of the words above the 2,000 word level in the vocabulary test in Nation (1990). It's old, it's based on a compilation of pre-WW2, non-computerized word counts totaling about 18 million written words. As published, it's not in frequency order, but frequency ranks are given for each word.

The University Word List

The University Word List (UWL)(in Nation, 1990) is a list of academic vocabulary composed of about 800 words. It's designed for students who plan to study in an English-language college or university. Essentially, it's the most common 800 words in academic texts, excluding the 2,000 words of the GSL. This list is structurally linked to the GSL. A student who studies the GSL, followed by the UWL, will find no repetition of words. The list is divided into 11 parts. Part one has the greatest frequency and range, part 2 next, etc. This list is also a component of the Vocabprofile program.

The Brown Corpus

The Brown Corpus (Francis and Kucera, 1982) is the earliest computerized study of English vocabulary. It is an analysis of 1 million words published in the United States in 1961. It's also kind of old, but it's more consistent in it's definition of "word" (as a lemma) than the earlier lists. The 1982 publication, which includes both alphabetical and frequency order lists of the words, is a very useful resource.

The LOB Corpus

The LOB Corpus (Hofland and Johansson, 1982) is a study of 1 million words of British text published in 1961. It was designed to be a British counterpart to the Brown corpus.

The Cambridge English Lexicon

The Cambridge English Lexicon (CEL) (Hindmarsh, 1980) is a list of 4470 words, prepared with reference to the GSL, Thorndike and Lorge, Brown, other sources, and the author's experience as an ESL teacher and material developer. Each item is graded from 1 to 5. The most useful aspect of the list is that the different meanings of the words are also graded on the same scale. Only the CEL and the GSL give separate information on the different meanings of common words (though, of course, dictionaries do also). The GSL gives actual frequency numbers for the different meanings, but the age of the data and the fact that it was gathered by hand may make the CEL a more reliable source for an indication of the relative importance to students of different meanings of words. The grading in the CEL is not based solely on frequency.

Modern Corpora

These days, much is heard about corpora from dictionary publishers, who all boast about the enormous corpora that their learner dictionaries are based on. The British publishers are particularly enthusiastic about this, using either the CoBuild corpus or the British National Corpus (BNC) as a source of lexicographic information. Both of these corpora contain more than 100 million words. Limited access to them is possible through the internet, see the links on the Collocations Homepage listed below. Depending on your purpose, it may be more useful to access these corpora in pre-digested form through the dictionaries based on them. A lemmatized frequency list of the BNC has been prepared by Adam Kilgarriff and is available for FTP.

Vocabprofile

Vocabprofile is a freeware program for PCs that will compare a given text with any properly formatted list. Three lists can be done at a time. The output will report what percent of the words in the text are on each of the lists. It will also print the text with the words marked to indicate which list they are on, or if they aren't on a list. Vocabprofile is available for FTP at the URL below. The three lists that come with the program are the first 1,000 words of the GSL, the second 1,000 words of the GSL and the UWL.

Concluding Remarks

None of these resources is ideal. Thorndike and Lorge and the GSL are old, old enough that the English of today surely differs significantly. However, the core vocabulary of English changes more slowly, so at the frequency level of the first 2,000 words this may be less of a problem. The GSL offers some advantages as a standard. It was specifically designed as a teaching vocabulary list. It has a long history of use, both in teaching materials and in second language acquisition research. A program to compare it with a given text is readily available. Of the lists above, only the CEL was also compiled for the purpose of facilitating the creation of teaching materials. It's more modern than the GSL, but appears to have had less impact. It is not conveniently available for computerized text comparison.

The Brown Corpus, the LOB Corpus and the lemmatized list from the BNC are useful because they give the lists in frequency order. This allows a population of words to be defined much more precisely, and individual words to be compared with each other. But these lists were prepared for linguistic research, not teachers. They're lists of lemmas, which means that words are listed more than once if they can act as more than one part of speech. Some derived forms are also considered as separate lemmas, such as comparative and superlative forms of adjectives. These factors affect both the frequency rankings of words and the number of words that appear on a list. In other words, a list of 1,000 words taken from the GSL or CEL would contain more than 1,000 lemmas. These corpus-based lists need substantial adjustment to make them appropriate as vocabulary standards. These adjustments have already been made to the GSL and CEL.

An author of EFL material has many vocabulary options available. I hope this discussion of resources is useful and that the bibliography and the internet sites below will be helpful in finding the items that will serve your specific needs.

Word Wide Web URLs

Adam Kilgarriff
http://www.itri.brighton.ac.uk/~Adam.Kilgarriff/
Links to his lemmatized, frequency order version of the BNC are here.

John Higgins
http://www.stir.ac.uk/epd/celt/staff/higdox/listers.htm
Here you can find Vocabprofile as well as links to other programs.

Collocations Homepage
http://www.ed.uiuc.edu/students/jc-lai/Fall95/
Jennifer Lai has collected links to corpora and other lexical resources here.

John Bauman
http://plaza3.mbn.or.jp/~bauman
This article, the UWL, the GSL, and some other resources are here.

Bibliography

Francis, W.N. and Kucera, H. (1982). Frequency Analysis of English Usage. Houghton Mifflin, Boston

Hindmarsh, R. (1980). Cambridge English Lexicon. Cambridge University Press, Cambridge

Hofland, K. and Johansson, S. (1982). Word Frequencies in British and American English. NAVF, Bergen

Nation, I.S.P. (1990). Teaching and Learning Vocabulary. Newbury House, New York

Thorndike, E.L. and Lorge, I. (1944). The teacher's Word Book of 30,000 Words. Teachers College, Columbia University, New York

West, M. (1953). A General Service List of English Words. Longman, London

FL 024860

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
Educational Resources Information Center (ERIC)

ERIC

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: Vocabulary Resources for Material Writers

Author(s): JOHN BAUMAN

Corporate Source: the Material Writers Newsletter Vol. IV no. 3

Publication Date: October 1996

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic/optical media, and sold through the ERIC Document Reproduction Service (EDRS) or other ERIC vendors. Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following two options and sign at the bottom of the page.

☒ Check here **For Level 1 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical) ***and*** paper copy.	The sample sticker shown below will be affixed to all **Level 1** documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) Level 1	The sample sticker shown below will be affixed to all **Level 2** documents PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN OTHER THAN PAPER COPY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) **Level 2**	□ Check here **For Level 2 Release:** Permitting reproduction in microfiche (4" x 6" film) or other ERIC archival media (e.g., electronic or optical), but ***not*** in paper copy.

Documents will be processed as indicated provided reproduction quality permits. If permission to reproduce is granted, but neither box is checked, documents will be processed at **Level 1**.

"I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic/optical media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries."

Sign here→ please

Signature: John Ba

Printed Name/Position/Title: John Bauman Instructor

Organization/Address:
TEMPLE UNIVERSITY Japan
2-8-12 Minami Azabu
MINATO-KU TOKYO JAPAN 106

Telephone: 022-218-9303

FAX: 022-218-9303

E-Mail Address: jbauman@aa.mbn.or.jp

Date: OCT. 21, 1997

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, *or*, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:

Address:

Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:

Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to:

~~ERIC Processing and Reference Facility~~
~~1100 West Street, 2d Floor~~
~~Laurel, Maryland 20707-3598~~

~~Telephone: 301-497-4080~~
~~Toll Free: 800-799-3742~~
~~FAX: 301-953-0263~~
~~e-mail: ericfac@inet.ed.gov~~
~~WWW: http://ericfac.piccard.csc.com~~

February 11, 1998

Kathleen Marcos
Acquisitions Coordinator
ERIC CAL

Dear Ms. Marcos,

I received your letter of January 6 regarding my contribution of the article

Vocabulary Resources for Material Writers.

I am the sole copyright holder. Please feel free to go ahead with processing the materials.

Sincerely,

John Bauman
Temple University Japan
2-8-12 Minami-Azabu
Tokyo 106
Japan